Genome analysis

BaiHui: cross-species brain-specific network built with hundreds of hand-curated datasets

Hong-Dong Li1,2, Tianjian Bai2, Erin Sandford3, Margit Burmeister2,3 and Yuanfang Guan2,*

1Center for Bioinformatics, School of Information Science and Engineering, Central South University, Changsha 410083, People’s Republic of China, 2Department of Computational Medicine and Bioinformatics and 3Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Functional gene networks, representing how likely two genes work in the same biological process, are important models for studying gene interactions in complex tissues. However, a limitation of the current network-building scheme is the lack of leveraging evidence from multiple model organisms as well as the lack of expert curation and quality control of the input genomic data.

Results: Here, we present BaiHui, a brain-specific functional gene network built by probabilistically integrating expertly-hand-curated (by reading original publications) heterogeneous and multi-species genomic data in human, mouse and rat brains. To facilitate the use of this network, we deployed a web server through which users can query their genes of interest, visualize the network, gain functional insight from enrichment analysis and download network data. We also illustrated how this network could be used to generate testable hypotheses on disease gene prioritization of brain disorders.

Availability and implementation: BaiHui is freely available at: http://guanlab.ccmb.med.umich.edu/BaiHui/.

Contact: gyuanfan@umich.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Modeling gene networks is an important approach to studying gene interactions and functions. The accumulation of genomic data has enabled the construction of accurate gene networks by integrating multi-line evidences of functional relationships between genes (Greene et al., 2015; Guan et al., 2008; Li et al., 2018). An essential application of gene networks is the prediction of disease-related genes. For example, by mining a mouse’s global (i.e. non-tissue-specific) gene network, we identified Timp2 and Abcg8 as novel bone-mineral density associated genes (Guan et al., 2010). Similarly, Recla et al. discovered Hydin to be a novel thermal pain gene (Recla et al., 2014). However, compared to global networks, tissue-specific networks (i.e. networks built using only one tissue’s data) can more accurately capture gene interactions and better assist in understanding diseases (Greene et al., 2015; Guan et al., 2012; Hu et al., 2010).

Yet, current tissue-specific network modeling approaches, including ours (Guan et al., 2012), are crude. Tissue-specificity is typically achieved by feeding gene expression data into the model without scrutinizing the input genomic data. This is problematic, as the gene expression data deposited in public repositories is dominated by cancers (cell lines) and the resulting networks is biased due to the excess of functional interactions present in cancers. Furthermore, the deposited genomic data are often generic, lacking in a hand-curation of tissue types. Finally, even though the experimental discovery process of gene functions and disease genes often involves multiple mammalian model species, current networks are...
We addressed the above limitations by intensively hand-curating thousands of non-cancerous, brain-specific gene expression datasets across humans, mice and rats using abstract reading. We then used Bayesian approaches to integrate these data to build a gene network, called BaiHui. This network eventually turned out to be accurate in discovering brain disease genes. We deployed the BaiHui web server, providing functions for users to query their genes of interest and to visualize the resulting local networks.

2 Materials and methods

Curation criteria of brain-specific gene expression data. We downloaded micro array gene expression data from the Gene Expression Omnibus via NCBI. After processing using a unified pipeline (Supplementary Text S1), 1304, 1215 and 320 micro array datasets were obtained for humans, mice and rats, respectively. A group of neurologists then hand-curated the 2839 total datasets by reading the original publication for their tissue source, species and health/dis-ease status and identifying a subset of 213 non-cancerous brain datasets (Supplementary Table S1). Every sample in each dataset was used.

Feature construction. Four types of features (Fig. 1) were constructed (detailed in Supplementary Text S1). Briefly, they are (i) Fisher’s z-transformed Pearson correlation coefficients (PCC) from the curated gene expression data; z-transformations make the distribution of PCC approximately normally distributed and comparable across multiple datasets, (ii) binary protein-protein interaction data from multiple public databases, (iii) protein-docking scores and (iv) phenotype sharing between gene pairs.

Network construction. By integrating the gold standard of functionally related gene pairs (Supplementary Text S1) and the four types of genomic data (Fig. 1A), a brain-specific gene network (BaiHui) was constructed using the established Bayesian network framework (Guan et al., 2008; Pe’er et al., 2006) (see details in Supplementary Text S2). In this network, each node is a gene and each edge represents the co-function probability (CFP) ranging from 0 to 1.

3 Results and applications

The network is accurate with a mean cross-validated AUC of 0.78 ± 0.006 (Fig. 1A). The network contains the CFP of all possible pairs of 21 000 genes. For query efficiency, for each gene, we identified its top 25 neighbors ranked by CFP (the neighbors with the same score as the 25th neighbor were also included). The local network of these top neighbors is stored in a MySQL database, and can be displayed online using the popular d3 javascript package. To gain biological insight into the local network, we implemented GO enrichment using the tool GOTermFinder and provided significantly enriched GO biological processes. Disease/trait enrichment was also implemented (Supplementary Text S3). A comparison of BaiHui to other types of networks is given (Supplementary Text S4). The BaiHui server is simple to use, with its usage detailed in Supplementary Text S5. By querying a gene, its local networks and enrichment results will all be shown on a single page, where a link is also provided for users to download the network data. Using the SNCA gene as an example, its network along with GO enrichment is illustrated (Fig. 1B).

An essential application of BaiHui is to prioritize research on disease genes, which could be both tedious and difficult without it. We illustrated this in two diseases: Parkinson’s disease (PD) and Alzheimer’s disease (AD). For PD, we queried the network (Fig. 1B) of a known disease gene SNCA. Firstly, we noted that the network is most enriched in GO terms concerning synaptic transmission and signaling, which were reportedly related to PD (Zhai et al., 2018). Secondly, motivated by the hypothesis that genes most connected to disease genes are most likely to be a possible disease gene, we looked at the top neighbors of SNCA. Of interest, the top two are SNAP25 and STX1A, which have been shown to be associated with PD (Bereczki et al., 2017; Borrageiro et al., 2018). We presented analysis of another PD gene PARK2 in Supplementary Figure S1. For AD, we queried BACE1, a known AD-related gene. Its network is most enriched in blood coagulation (Fig. 1C), which are known biological processes associated with AD. We also found that its top two ranked neighbors (PSEN1 and CTSD) have been implicated in AD (Lanoiselee et al., 2017; Schaur et al., 2011). The network analysis of a second AD gene APOE is given in Supplementary Figure S2. We also provided the analysis of two more example genes, SMN2 (Supplementary Fig. S3) and RBFOX1 (Supplementary Fig. S4) (Voinneau et al., 2011). These results support that BaiHui could be a valuable resource for prioritizing disease genes.

4 Conclusions

We have built a brain-specific functional gene network and developed the BaiHui web server as a unique resource allowing users to query networks of their genes of interest. As illustrated in this work, this network can be used to generate hypotheses about disease candidate genes of brain disorders including PD and AD. A salient advantage of this network is the intensive curation of non-cancerous data. Because the hand-curation of data is highly intensive work and time-consuming, we only included microarray data, which is one limitation of this network. In the future, we will continue to adapt the data gathering (such as incorporating RNA-seq and protein expression data), curation and processing to improve the network.
Funding
This work was supported by the National Science Foundation of China [61702556 to H.D.L.] and the startup funding [202041004 to H.D.L.].

Conflict of Interest: none declared.

References